University of California, Los Angeles
Department of Statistics

Statistics 13 **Instructor: Nicolas Christou**

Homework 1 - Solutions

EXERCISE 1:

a. Both lists have the same sample mean. You don't need to compute it. Let's examine the $\sum_{i=1}^n (x_i - \bar{x})^2$ for both lists.
List P: $\sum_{i=1}^n (x_i - \bar{x})^2 = 2(1-\bar{x})^2 + 2(2-\bar{x})^2 + \cdots + 2(1000-\bar{x})^2 = 2A$.
List Q: $\sum_{i=1}^n (x_i - \bar{x})^2 = 3(1-\bar{x})^2 + 3(2-\bar{x})^2 + \cdots + 3(1000-\bar{x})^2 = 3A$.
The sample variance of list P is: $s_P^2 = \frac{2A}{1999}$ and the sample variane of Q is $s_Q^2 = \frac{3A}{2999}$. Therefore the standard deviation of Q is smaller but just by a little. Note that if these 2 lists were populations the two standard deviations would have been equal.

b. Both lists have the same sample mean (200). It is easy to see that the values of list P are more dispersed. Therefore list Q has smaller standard deviation.

c. For both lists the sample mean is 20. For list P we have: $\sum_{i=1}^n (x_i - \bar{x})^2 = 100(18-20)^2 + 100(19-20)^2 + 100(20-20)^2 + 100(21-20)^2 + 100(22-20)^2 = 100(4+1+0+1+4) = 1000$. For list Q we have: $\sum_{i=1}^n (x_i - \bar{x})^2 = 150(18-20)^2 + 200(19-20)^2 + 150(22-20)^2 = 150(4+0+4) = 1200$.
It follows that list P will have the smaller standard deviation.

EXERCISE 2:

New sample mean and sample variance in terms of $\bar{x}$ and s^2 when a constant a is added to the original observations or when the original observations are multiplied by a constant a.

a. The new observations will be $y_1 = x_1 + a, y_2 = x_2 + a, \ldots, y_n = x_n + a$. The new sample mean is $\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{\sum_{i=1}^n (x_i + a)}{n} = \bar{x} + a$. The new variance is $s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n (x_i + a - \bar{x} - a)^2}{n-1} = s_x^2$.

b. The new observations will be $y_1 = ax_1, y_2 = ax_2, \ldots, y_n = ax_n$. The new sample mean is $\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{\sum_{i=1}^n ax_i}{n} = a\bar{x}$. The new variance is $s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n (ax_i - a\bar{x})^2}{n-1} = a^2 s_x^2$.

EXERCISE 3:

The formula that converts Farenheit degrees into Celcius degrees is: $C = (F-32)\frac{5}{9}$, where C is celcius degrees and F is Farenheit degrees. We know that the average temperature in Los Angeles is 85 Farenheit degrees with standard deviation 10 Farenheit degrees. $C = \frac{5}{9}F - \frac{160}{9}$. Therefore the average temperature in Celcius degrees is $\frac{5}{9}85 - \frac{160}{9} = 29.44$ Celcius degrees. The standard deviation in celcius degrees is $\frac{5}{9}10 = 5.56$ Celcius degrees.

EXERCISE 4:

a. The sum of the 101 observations is $101(240.0) = 24240$. The new sum is:
New sum$=24240 - 230 + 200 - 250 + 280 = 24240$. Therefore the new mean is going to be the same as the old one: $\bar{x} = 240.0$.

b. The sample variance is given by $s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$, therefore we can compute:
$\sum_{i=1}^n (x_i - \bar{x})^2 = (n-1)s^2 = (101-1)(25.88)^2 = 66977.44$. We must subtract from this calculation the quantity $(230-240)^2 + (250-240)^2 = 200$, and add the quantity $(200-240)^2 + (280-240)^2 = 3200$. The new $\sum_{i=1}^n (x_i - \bar{x})^2 = 66977.44 - 200 + 3200 = 69977.44$. And finally the new sample variance is $s^2 = \frac{69977.44}{100} = 699.7744$, and the sample standard deviation is $s = \sqrt{699.7744} \Rightarrow s = 26.45$.

EXERCISE 5:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) =$$

$$\frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2] =$$

$$\frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - 2\frac{(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}x_i)}{n} + n(\frac{\sum_{i=1}^{n}x_i}{n})^2] =$$

$$\frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - 2\frac{(\sum_{i=1}^{n}x_i)^2}{n} + \frac{(\sum_{i=1}^{n}x_i)^2}{n}] =$$

$$\frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{(\sum_{i=1}^{n}x_i)^2}{n}] =$$

$$\frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - \frac{n(\sum_{i=1}^{n}x_i)^2}{n^2}] =$$

$$\frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - n(\frac{\sum_{i=1}^{n}x_i}{n})^2] =$$

$$\frac{1}{n-1}[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2]$$

EXERCISE 6:
The sample covariance is given by $\text{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$.

$$\begin{aligned}
\text{cov}(x,y) &= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_iy_i - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}(x_iy_i - \bar{y}x_i - \bar{x}y_i - \bar{x}\bar{y}) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i - \bar{y}\sum_{i=1}^{n}x_i - \bar{x}\sum_{i=1}^{n}y_i + n\bar{x}\bar{y}\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}\right) = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_iy_i - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)\right].
\end{aligned}$$

Note: $\bar{y}\sum_{i=1}^{n}x_i = \bar{x}\sum_{i=1}^{n}y_i = n\bar{x}\bar{y}$.

EXERCISE 7:
From exercise 3b we find that the new standard deviations will be $c_1sd(x)$ and $c_2sd(y)$. Also the covariance will be $c_1c_2cov(x,y)$ and therefore the new correlation will no change because

$$r = \frac{c_1c_2cov(x,y)}{c_1sd(x)c_2sd(y)} = cov(x,y).$$

EXERCISE 8:
To plot the empirical cumulative distribution function we use the formula $F_n(x) = \frac{\# x_i's \leq x}{n}$. Make sure you have the numbers sorted from smallest to largest. We get the following table:

x	$F_n(x)$
1.5	$\frac{5}{20}$
3.6	$\frac{6}{20}$
5.8	$\frac{7}{20}$
10.5	$\frac{9}{20}$
13.0	$\frac{10}{20}$
16.0	$\frac{13}{20}$
20.0	$\frac{14}{20}$
25.0	$\frac{15}{20}$
30.0	$\frac{16}{20}$
31.0	$\frac{17}{20}$
50.0	$\frac{18}{20}$
60.0	$\frac{19}{20}$
61.0	$\frac{20}{20}$

The plot is given below with the `R` commands:

```
x <- c(1.5, 1.5, 1.5, 1.5, 1.5, 3.6, 5.8, 10.5, 10.5, 13.0, 16.0, 16.0, 16.0,
      20.0, 25.0, 30.0, 31.0, 50.0, 60.0, 61.0)

y <- cdf(x)

y <- ecdf(x)

plot(y, do.points=FALSE, verticals=TRUE,
     main="Empirical cumulative distribution function")
```